Tree Genetics & Genomes (2012) 8:549–557
DOI 10.1007/s11295-012-0494-7

REVIEW

Uniform standards for genome databases in forest and fruit trees

J. L. Wegrzyn · D. Main · B. Figueroa · M. Choi · J. Yu · D. B. Neale · S. Jung · T. Lee · M. Stanton · P. Zheng · S. Ficklin · I. Cho · C. Peace · K. Evans · G. Volk · N. Oraguzie · C. Chen · M. Olmstead · G. Gmitter Jr. · A. G. Abbott

Received: 4 April 2011 / Accepted: 16 February 2012 / Published online: 27 March 2012
© The Author(s) 2012. This article is published with open access at Springerlink.com

Abstract TreeGenes and tree fruit Genome Database Resources serve the international forestry and fruit tree genomics research communities, respectively. These databases hold similar sequence data and provide resources for the submission and recovery of this information in order to enable comparative genomics research. Large-scale genotype and phenotype projects have recently spawned the development of independent tools and interfaces within these repositories to deliver information to both geneticists and breeders. The increase in next generation sequencing projects has increased the amount of data as well as the scale of analysis that can be performed. These two repositories are now working towards a similar goal of archiving the diverse, independent data sets generated from genotype/phenotype experiments. This is achieved through focused development on data input standards (templates), pipelines for the storage and automated curation, and consistent annotation efforts through the application of widely accepted ontologies to improve the extraction and exchange of the data for comparative analysis. Efforts towards standardization are not limited to genotype/phenotype experiments but are also being applied to other data types to improve gene prediction and annotation for de novo sequencing projects. The resources developed towards these goals represent the first large-scale coordinated effort in plant

Communicated by T. Drudge

A contribution to the special issue "The genomes of the giants: a walk through the forest of tree genomes".

J. L. Wegrzyn (✉) · B. Figueroa · M. Choi · J. Yu · D. B. Neale
Department of Plant Sciences, University of California at Davis,
Mail Stop 4,
Davis, CA 95616, USA
e-mail: jlwegrzyn@ucdavis.edu

D. Main (✉) · S. Jung · T. Lee · P. Zheng · S. Ficklin · C. Peace · K. Evans · N. Oraguzie
Department of Horticulture and Landscape Architecture,
Washington State University,
PO Box 646414, Pullman, WA 99164, USA
e-mail: dorrie@wsu.edu

I. Cho
Department of Computer Science,
Saginaw Valley State University,
Saginaw, MI 48710, USA

M. Stanton
Clemson University Genomics Institute, Clemson University,
Clemson, SC 29634, USA

G. Volk
National Center for Genetic Resources Preservation, USDA/ARS,
Fort Collins, CO 80521, USA

M. Olmstead
Horticultural Sciences Department–IFAS, University of Florida,
Gainesville, FL 32611, USA

C. Chen · G. Gmitter Jr.
Citrus Research and Education Center, University of Florida,
Lake Alfred, FL 33850, USA

A. G. Abbott
Department of Genetics and Biochemistry, Clemson University,
Clemson, SC 29634, USA

databases to add informatics value to diverse genotype/phenotype experiments.

Keywords Phenotype · Ontologies · Genome database · Genotype · Web services · Genome annotation

Introduction to databases

Genome databases are essential resources for experimental and computational biologists. Genome-related databases can be broken into two major groups: generalized and specialized (domain) databases. Generalized repositories include the Genbank/DDJB/EMBL databases of nucleic acids sequences and the PIR/SwissProt/PDB protein sequence databases. These resources capture and deliver information on specific classes of molecules, without any phylogenetic or functional restrictions. In contrast, the specialized databases are more limited in scope and are organized around a specific model organism or biological function (tissue type or protein family). Comprehensively, collected sequence data provide essential genomic resources for accelerating molecular understanding of biological properties and for furthering the application of this knowledge.

Often, both types of databases contain a mixture of data from genome projects and supporting studies from the broader scientific community. Although the contributions of the community might lack data consistency and breadth of coverage, these possible deficiencies are offset by the greater expertise behind the individual contributions (from years of focused research). The construction of well-designed bioinformatics platforms and databases allows users to take advantage of diverse data sets and provides a foundation for comparative genomics. Promotion of comparative genomics among model and applied plants allows researchers to grasp the biological properties of each species and to accelerate gene discovery and analyses. Recent progress in plant genomics has discovered and isolated many important genes with functions that increase yield, quality, and tolerance to various environmental stresses.

The TreeGenes database (http://dendrome.ucdavis.edu/TreeGenes) is home to comprehensive genomic data on conifers and other forest tree species (Wegrzyn et al. 2008). It serves as the primary international resource on genetic maps, resequencing, genotyping, and phenotyping in forest trees (Fig. 1). TreeGenes provides custom informatics tools to manage the flood of information resulting from high-throughput genomics projects from sample collection to downstream analysis. This resource is further enhanced with systems that are well connected with federated databases, automated data flows, machine learning analysis, standardized annotations, sequence search tools, and quality control processes. A sample tracking system now sits at the forefront of most large-scale projects. Barcode identifiers assigned to individual trees during sample collection are maintained in the database to identify an individual through DNA extraction, resequencing, genotyping, and phenotyping. Emerging technologies have been applied to integrate a solution for high-throughput SNP discovery in non-model organisms. The Pine Sequence Alignment and SNP Identification Pipeline identifies SNPs from both Sanger and 454 sequencing that reflect true genetic variation (Wegrzyn et al. 2009). The database itself contains ten curated modules that support the storage of data and provide the foundation for web-based searches and visualization tools. DiversiTree (http://dendrome.ucdavis.edu/DiversiTree/), an extensive user-friendly desktop-style interface, queries the TreeGenes database and is designed for bulk data retrieval. It provides the community with access to a multitude of data types including ESTs, primers, trace files, SNPs, individual tree data, genotypes, and phenotypes. DiversiTree also connects directly to the Forest Tree Genetic Stock Center (http://dendrome.ucdavis.edu/ftgsc/) where users can order specific DNA based on sample, sequence, or maker data. The combined resources serve as a powerful knowledge environment for genotype–phenotype information resulting from a multitude of large-scale genomics projects.

Tree fruit Genome Database Resources (tfGDR, www.tfgdr.org) is a centralized worldwide repository of genomics and genetics data for Rosaceae (www.rosaceae.org) and *Citrus* (www.citrusgenome.org) species (Jung et al. 2008). TfGDR provides a common bioinformational infrastructure for collecting, integrating and translating the large and diverse amounts of structural, functional, and comparative genomics data into a knowledgebase serving gene discovery, marker-trait identification, and genomics-assisted breeding for tree fruit and related crops (Fig. 1). The database is implemented in Tripal (Ficklin 2011) an open-source framework that blends the power of a web content management system and Generic Model Organism Database (GMOD) chado (www.gmod.org). Tripal is under active development by several collaborating institutions and is part of the GMOD suite of tools. TfGDR serves as a primary resource for access to the genome sequences of apple, peach, mandarin orange, sweet orange, and strawberry and is being further expanded to include the cacao (www.cacaogenomedb.org) and blueberry genome sequences (www.vaccinium.org). Genome sequence, transcriptome, and mapping data are available using the widely used GMOD tools GBrowse (Stein et al. 2002; Donlin 2009), CMap (Youens-Clark et al. 2009), and GBrowse-Syn (Mckay et al. 2010). Custom interfaces provide easy-to-use, clear routes to specific data, and associated data types regardless of database point of entry. Custom computational analysis pipelines are available for functional annotation of genes, transcripts, and markers. Researchers can download, browse or search these datasets using standard interfaces. All sequence datasets in tfGDR are available for searching via standalone NCBI Basic Local Alignment Search Tool (BLAST) and custom batch

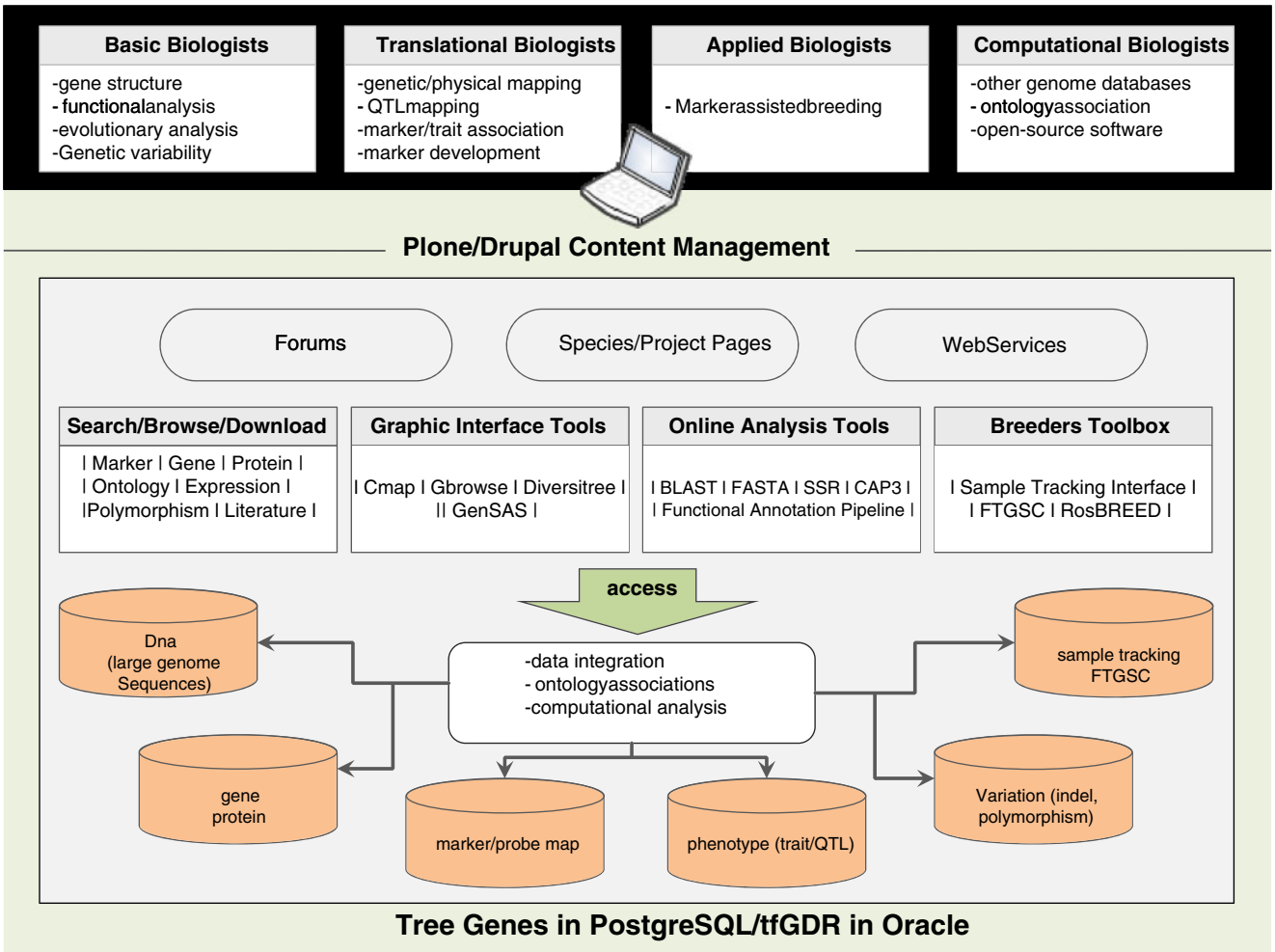

Fig. 1 This schematic depicts the overall organization of the TreeGenes and tfGDR database systems. The *top boxes* represent the diverse categories of users that access the system and a few of their research objectives. Despite using different platforms for both the relational database system and the content management software, the sequence data types and resources developed are nearly identical providing opportunity for collaboration in standards and tool development

BLAST tools. Other standalone online tools include sequence assembly and microsatellite marker identification. TfGDR provides a home for community projects through its secure, private Drupal content management interface and aids research collaboration and communication through custom community mailing lists and searchable archives.

In order to provide an effective platform for genomics research, a public database must make data available in a user-friendly, functional manner. When considering the utility of a genome database, four key attributes come to mind: (1) Flexibility allows a system to adapt quickly to changing data types, both in terms of quantity and type. Next generation sequencing, high-throughput genotyping/phenotyping, and large-scale expression studies are common data types. Achieving flexibility involves technical foresight in templates for interfaces and back-end database organization. (2) Standards adhering to community consensus as well as applying formalized nomenclature are important to the organism focus. Describing the function of genes and phenotypes through ontologies, as well as enforcing nomenclatures for genetic markers and novel genes is critical to support comparative analysis. (3) Integration allows for effective data type connections within the database, as well as among related databases. Some of these relationships are derived computationally, while others are manually curated. The strengths of these relationships also vary; some resources share data directly leading to exact mappings between their objects, while others report common attributes of their objects. (4) Interfaces require a great deal of thought and technical input to be both user-friendly and efficient. A repository that is difficult to navigate from a user's perspective will be quickly abandoned for other resources. Single, form-based searches as well as bulk downloads should be supported for nearly all genomic data types. Examples of how the primary tree fruit and forestry databases are working towards standards and integration between and within their communities are discussed here.

Standards for submission

The field of genomics is faced with the novel file formats, new sequencing technologies, and increasing amounts of data. Genome databases are asked to remain flexible and at the same time adhere to quickly changing community standards. Specialized databases often turn toward the generalized members of the broader genomics community, such as NCBI, for guidance and as the first step for integrating data. Many genetic data forms already have a place for submission within those generalized repositories, including: ESTs, mRNAs, SNPs, polypeptide sequences, and expression data (Table 1). Both TreeGenes and tfGDR regularly import these data types from public repositories and repackage them to provide comparative value to the research community. Some of these enhancements include aligning sequence to existing genomes, re-clustering within species groups, and organizing subsets of sequence data for bulk downloads. For example, TreeGenes and tfGDR provide a custom forest tree and tree fruit annotation pipelines that, in their specificity, improve the characterization of a given sequence. Attempts at collecting this raw sequence data outside or independently of generalized repositories would result in inconsistencies and redundancies downstream. In addition, most reputable publishing groups require the submission of all sequence-based data to the NCBI repository. Reviewers and editors are explicitly directed to check that each data type is submitted *in advance* of *review*. This widely accepted requirement has greatly increased the availability of data to genomics researchers. What is lacking in these required sequence submissions is a resource to capture the genotypic, phenotypic, and geographic information that permits much of the analysis in forestry and horticulture genomics. There is currently no resolve or requirement for formal submission of this type of data. This information is generally appended as inconsistently formatted supplemental text files that cannot be easily accessed, directly compared, or readily queried.

One role of the specialized database is to develop standards for the data types not included in the larger repositories. The standards for the submission of this information are for the first time aligned with the support of subject journals, such as *Tree Genetics and Genomes*. The data types in this category include genetic maps, genotype, phenotype, and environmental data. Genetic maps were the first and principal data type for the TreeGenes and tfGDR databases. They serve as the foundation for bridging genomic and phenotypic data. The value of a computational resource of genetic maps is measured by its ability to provide comparison. Packages, such as GMOD's Comparative Map Viewer (CMap), have provided the foundation for the 62 and 48 genetic maps currently available in the TreeGenes and tfGDR database, respectively. To encourage users to submit their mapping data, a template is provided to record markers and positions along linkage groups according to specific nomenclature requirements. This information is associated with a literature object and is uploaded, reviewed, and assigned a unique accession number. The standardized nomenclature of the markers allows for the generation of automated mappings in the database and comparative map sets within and between tree species. Similar to other sequence databases, the map associated with that accession can be released immediately to the public or held until publication. The accession number is intended for use in manuscripts as a method to uniquely link and identify the map set.

The plans for organization of genotype/phenotype data in TreeGenes/tfGDR developed through two independent USDA/NIFA projects: Marker-Assisted Breeding in Rosaceae (RosBREED) and the Conifer Translation Genomics Network (CTGN). These were very large-scale evaluations of genotypes and phenotypes across multiple species. Some integrated applications are publicly available to bring together phenotypes, genotypes, and resulting associations. NCBI has launched dbGaP (Zhang et al. 2008), a public database to archive genotype and clinical phenotype data from human studies. The Complex Trait Consortium has launched GeneNetwork (Wu

Table 1 Repositories for submission of data

Data type	tfGDR	TreeGenes
Nucleotide sequence data: mRNA, cDNA, BAC	NCBI Genbank submissions	NCBI Genbank Submissions
Single nucleotide polymorphisms (SNPs)	NCBI dbSNP submissions	NCBI dbSNP submissions
EST sequence data	NCBI dbEST submissions	NCBI dbEST submissions
Transcriptome assembly	NCBI TSA submissions	NCBI TSA submissions
Next generation sequencing reads	NCBI SRA submissions	NCBI SRA submissions
Protein sequence data	NCBI SwissProt/Trembl	NCBI SwissProt Trembl
Gene expression studies	NCBI GEO submissions	NCBI GEO submissions
Genetic maps	Direct submissions at GDR	Direct submissions TreeGenes
Genotype/phenotype	Direct submissions at GDR	Direct submissions at TreeGenes
Gene annotations	Submissions through GenSAS	Submissions through GenSAS

et al. 2004), a database for mouse genotype, classical phenotype, and gene expression phenotype data with tools for "per-trait" quantitative trait loci analysis. PhenomicDB is a multi-species solution for non-plant model organisms by Metalife (Groth et al. 2007). None of the existing tools currently handle plant data or geographic information which is vital for the increasing number of landscape genomics studies. Environmental variables such as GPS coordinates, elevation, soil type, and precipitation measures are critical component for many areas of plant comparative genomics. Within TreeGenes, a sample tracking system was developed for CTGN that provides protocols for needle/leaf collection and a mechanism for tracking barcoded tissue samples through the DNA extraction and genotyping process. This same system provides a flexible but well defined interface for individuals to submit location information, environmental descriptors, and phenotypic evaluations (Fig. 2). This interface can gather information from common garden experiments, established breeding plantations and large or small-scale tree samples across a landscape. Recently, the pipeline was expanded to accommodate submissions downstream through unique sample identifiers, tree descriptors, and phenotype values within the range of defined metrics. This information is stored and organized by institution so that studies using the same individual trees can be associated, thereby adding value (and providing for) future analysis. Similar to microarray studies in NCBI's Gene Expression Omnibus (Edgar et al. 2002), unique accession numbers are assigned to each study and include metadata on the design as well as all related values. Within tfGDR, a prototype breeding data management system has been developed for the WSU Apple Breeding Program. Pedigree, phenotype, and genotype data are uploaded from a template provided by the breeder and stored in the database. Interfaces have been developed to allow the breeder to browse or search this private data by project, site, cultivar, trait, or pedigree and view or download the data in community-agreed standard formats suitable for statistical analysis. Further work will connect this private data to tfGDR public data for the breeder to provide a comprehensive breeders toolbox with cross planning and seedling selection functionality as part of the RosBREED project. This breeding data management system will be expanded, like the TreeGenes sample tracking interface, to gather outside information and provide accessions for fruit tree genotype/phenotype studies.

The application of standards does not end with the submission phase. Once data are collected consistently, it must be stored, curated, and annotated in a way that adds value and provides for future extraction and comparative analysis. Ontologies provide a shared and controlled vocabulary that can be used to model the domain in terms of the types of object or concept and their properties and relationships. Ontology is more complex than systematics used for species classification because it involves multiple parents and the

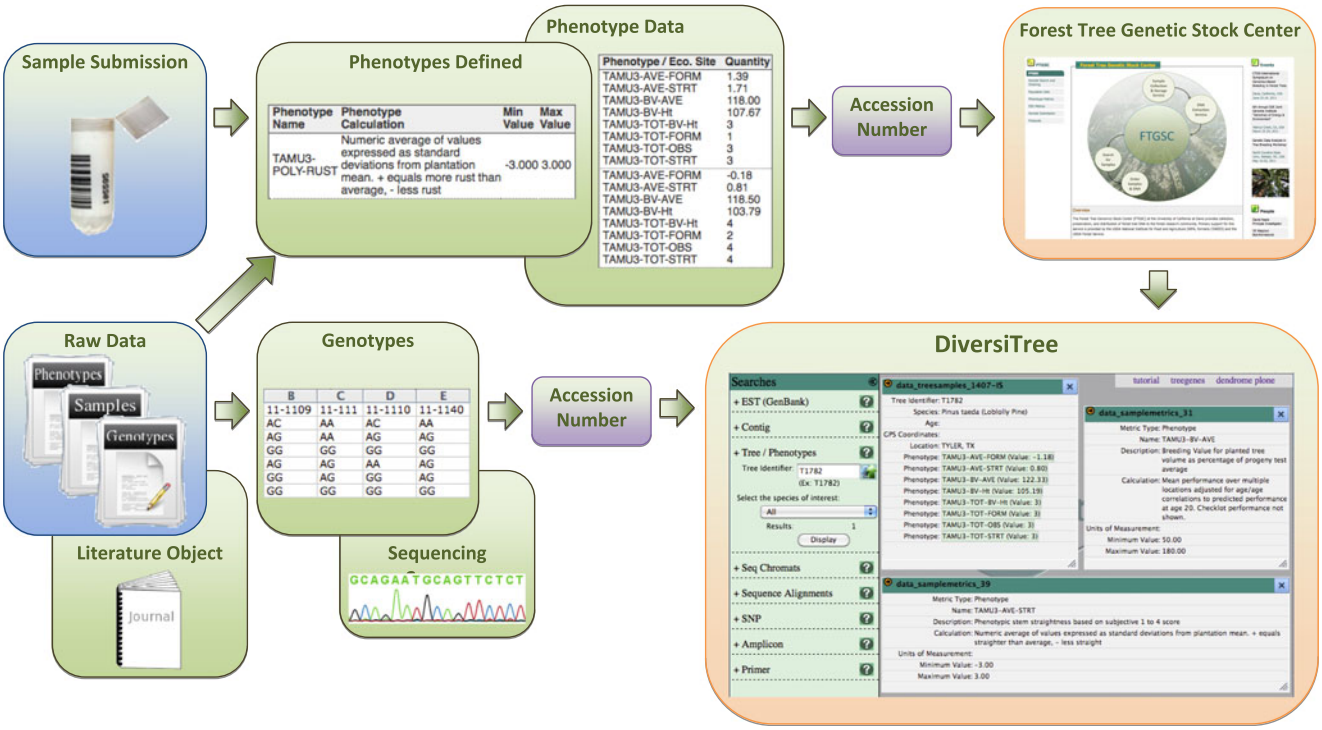

Fig. 2 Development of the sample tracking interface within the TreeGenes database provides a mechanism for submitting barcoded samples and associated each sample with phenotype data collected and genotype data received directly from sequencing centers. Expansion of this system now allows submission of phenotype–genotype data assigned to a unique identifier. This method provides a user the necessary templates to submit their data and obtain an accession number for publication purposes

opportunity for different relationships. While the structure of an ontology is a strict hierarchy, it is represented by a directed acyclic graph in which multiple types can have parents, with different relationships between them. Ontologies greatly enhance the value of databases by allowing users to query the different databases or subsections of a database using the same keywords and query strings. Gene Ontology (Ashburner et al. 2000) was one of the original tools and was designed around three categories that describe the function(s), process(es), and location(s) of a specific gene's expression. Plant ontology captures descriptions of organism structures and developmental stages as they relate specifically to plants (Jaiswal et al. 2005). The newer Environmental Ontology (EnvO) describes the habitat of a given organism or population. This descriptive vocabulary links the values of an organism to its environment. Phenotype, Attribute and Trait Ontology (PATO) is an ontology of phenotypic qualities, intended primarily for phenotype annotation and composite phenotypes. PATO is designed to be used in conjunction with ontologies of quality-bearing entities. An example of such an entity is a stem (from plant ontology), which could be the bearer of the quality "dark green" (PATO:0000328). Both TreeGenes and tfGDR maintain and

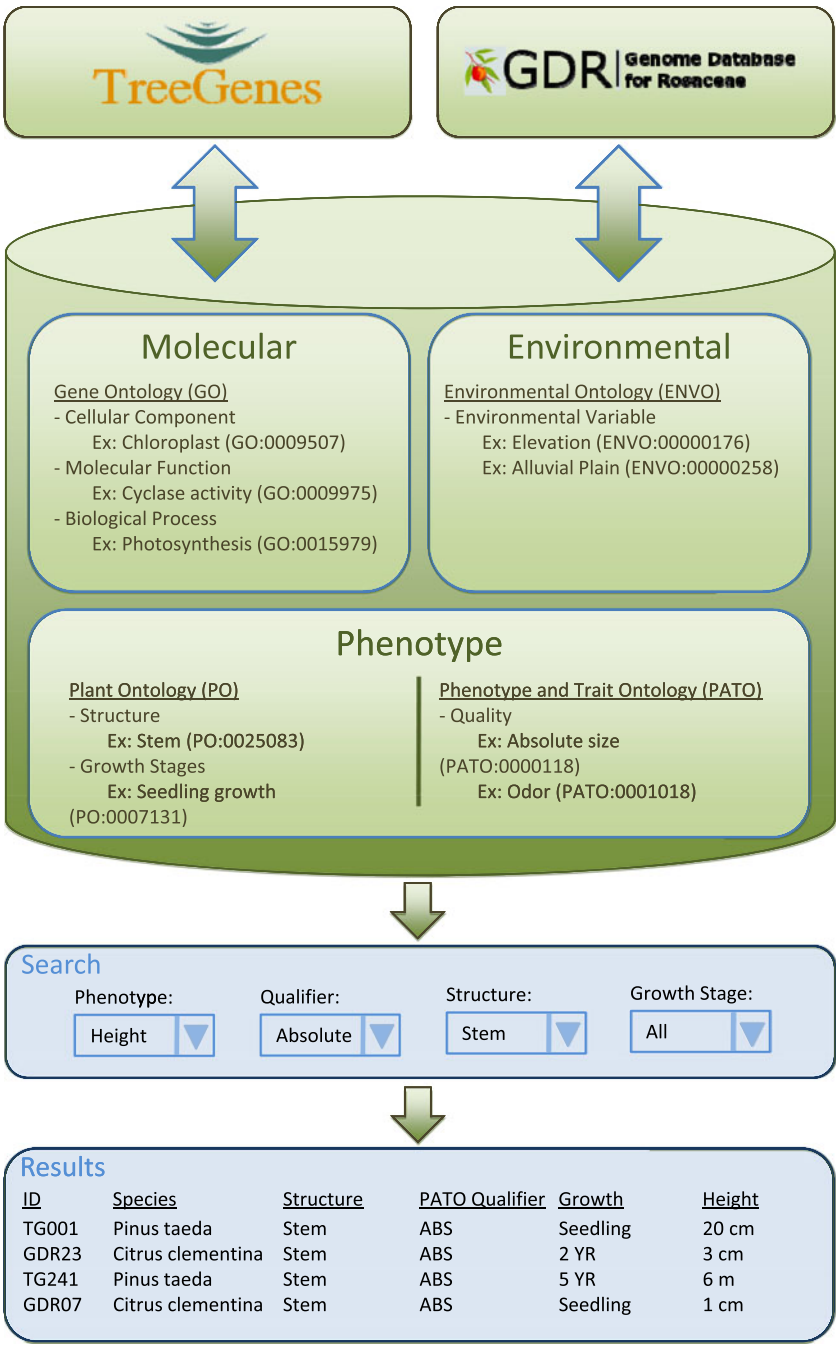

Fig. 3 Relevant consortium-guided ontologies are stored as local copies within the TreeGenes and tfGDR databases. The vocabularies are applied to genes, plant phenotypes, phenotype qualifiers, and environmental observations as part of the curation and annotation efforts that occur downstream from submission. Application of ontologies improves the ability to extract, exchange, and analyze the sequence data. Contribution of terms back to the ontology consortiums by the fruit tree and forestry databases is important to make species-specific vocabulary available

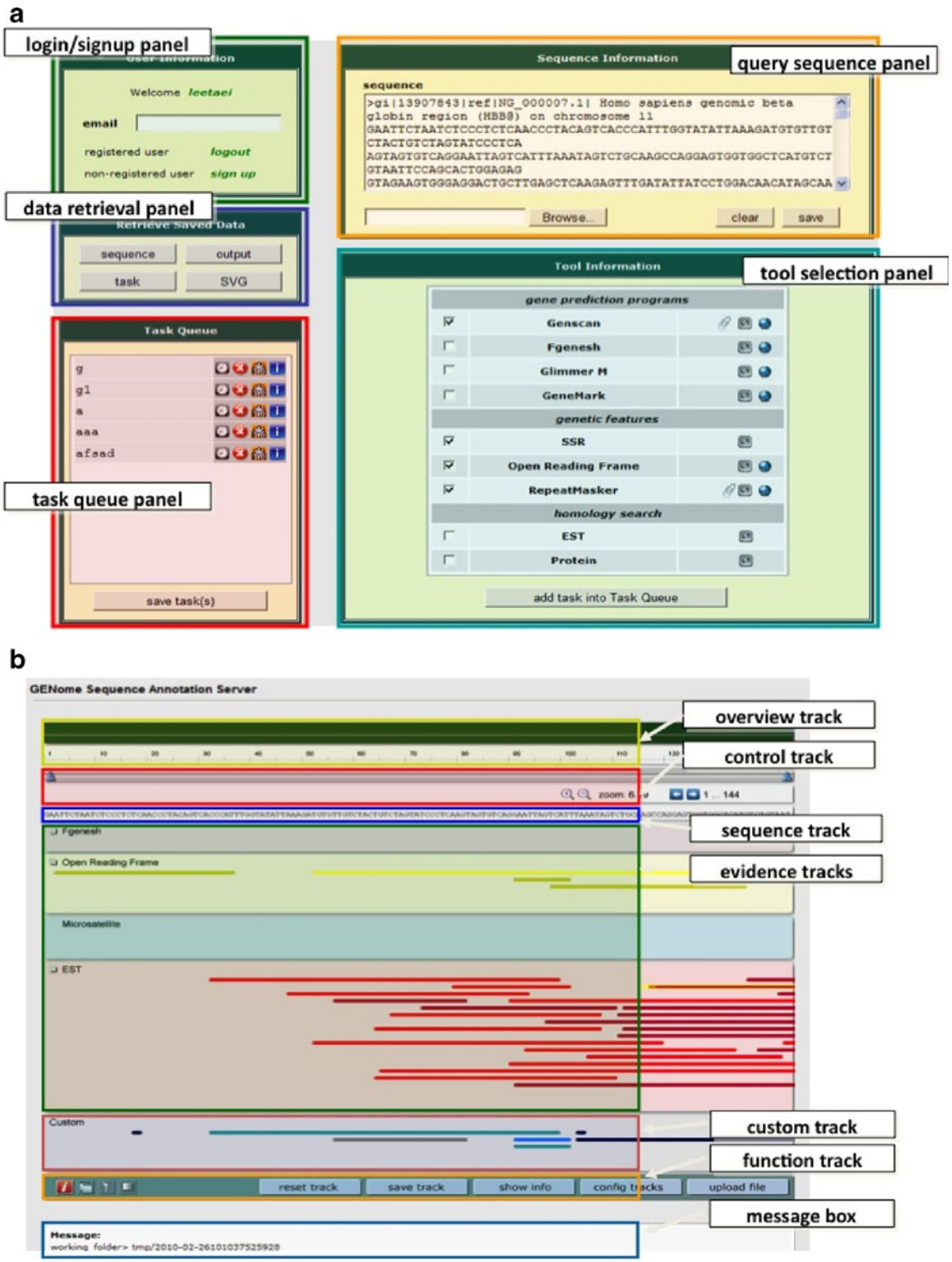

Fig. 4 Genome Sequence Annotation Server (GenSAS) is being developed to provide fruit tree and Rosaceae researchers with access to a collaborative online annotation tool. Through a dedicated secure web interface, users will have access to personalized individual or group analysis space for annotating their genome sequences. **a** Through a dedicated secure web interface, users have access to personalized individual or group analysis space for annotating their genome sequences. This includes the ability to upload and store sequences and analyses, identify and mask repeats, predict open reading frames and microsatellites, run gene prediction programs (or upload output from gene prediction programs), map transcripts from custom EST/cDNA databases and perform protein homology searches against the reference databases SwissProt and TREMBL, and custom protein databases. **b** The results of the analysis are displayed on tracks and users then select the evidence they wish to accept on a custom track. A core component of GenSAS final development will be full integration with GBrowse to provide a platform for comparative genomics among multiple species

update local copies of these repositories (Fig. 3). Gene Ontology and Plant Ontology have been well integrated into the functional annotation pipelines used by TreeGenes and tfGDR. Much work is yet needed to fully integrate PATO into phenotype submissions and encourage the community to work with these controls to improve the value of phenotypes available. It is the role of the specialized databases to represent the community and submit information back to the ontology consortiums to describe unique plant structures, phenotypes, and phenotype qualities.

Future directions: toward integration

In preparation for the sequencing of three conifer genomes (*Pinus taeda*, *Pinus lambertiana*, and *Pseudotsuga menziesii*), the primary comparative forestry and fruit tree databases have paired up to apply their vast resources to the challenge of annotating of the largest genomes sequenced. One of the first objectives under this project will be the integration of a web-based annotation tool known as the Genome Sequence Annotation Server (GenSAS) which is being developed to provide fruit tree and Rosaceae researchers access to a collaborative online annotation tool. Through a dedicated secure web interface, users will have access to personalized individual or group analysis space for annotating their genome sequences. This includes the ability to upload and store sequences and analyses, identify and mask repeats, predict open reading frames and microsatellites, run gene prediction programs (or upload output from gene prediction programs), map transcripts from custom EST/cDNA databases and perform protein homology searches against the reference databases SwissProt and TREMBL, and custom protein databases. The results of the analysis are displayed on tracks (Fig. 4) and users then select the evidence they wish to accept on a custom track.

This application will be further developed to serve as a tool for the community to curate gene/gene families and provide more comparative analysis across tree and other model plant species. This system will allow researchers to readily curate their gene/gene families of interest from an intuitive interface and take advantage of a comparative platform without an extensive bioinformatics background. A core component of GenSAS final development will be full integration with GBrowse. This web-based genome browser is a product of the GMOD project and has been widely accepted into the genomics community (Stein et al. 2002). The software can readily accept and exchange the GFF3 files that will be used both for the initial manual annotations and the modifications/additions that result from manual curations in GenSAS.

Integration for the purposes of genome annotation is just one area where comparative analysis is necessary. Both TreeGenes and tfGDR use different back-end platforms to house the data that are not directly comparable. In recent years, many databases have overcome this challenge through the application of web services. BioMart is a widely used platform for improving communication between databases sitting on different platforms (Smedley et al. 2009). BioMart enables researchers to perform advanced querying of genomics data sources through a single web interface. The power of the system comes from integrated querying of data sources regardless of their physical locations and database designs. Once these queries have been defined, they may be automated with its "scripting at the click of a button" functionality. These capabilities are extended by integration with various analysis and visualization software packages such as BioConductor, DAS, Galaxy, and Cytoscape. Comparative specialized plant databases, such as Gramene, have also applied this integrated web service functionality to give users access to data housed in other repositories. Both TreeGenes and tfGDR will utilize web services to provide queries to improve the exchange of data and therefore expand the comparative genomics potential of their respective systems.

TreeGenes and tfGDR share many similarities in the type of data, analysis, and informational resources they provide to their respective forest and tree fruit communities. Where they differ provides opportunity for synergistic collaboration. Where tfGDR already has access to several draft or completed tree fruit genome sequences, the two groups can work together to develop a common community-based gene curation system platform that meets the needs of both the forestry and the horticultural community, while eliminating redundancy of effort. Similarly, TreeGenes has more experience in providing users with access to genotype and phenotype data, an approach tfGDR has begun to leverage to their advantage. Together, the two databases can work toward uniform data standards and access that will benefit both research communities and provide greater opportunities in comparative genomics.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author (s) and the source are credited.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM et al (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25(1):25–29

Donlin MJ (2009) Using the Generic Genome Browser (GBrowse). Curr Protoc Bioinformatics, Chapter 9, Unit 9.9

Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30(1):207–210

Ficklin S (2011) Tripal: a construction toolkit for online genomic databases. Plant and Animal Genome Conference, GMOD Workshop. San Diego

Groth P, Pavlova N, Kalev I, Tonov S, Georgiev G, Pohlenz HD et al (2007) PhenomicDB: a new cross-species genotype/phenotype resource. Nucleic Acids Res 35(Database issue):D696–D699

Jaiswal P, Avraham S, Ilic K, Kellogg EA, McCouch S, Pujar A et al (2005) Plant ontology (PO): a controlled vocabulary of plant structures and growth stages. Comp Funct Genomics 6(7–8):388–397

Jung S, Staton M, Lee T, Blenda A, Svancara R, Abbott A et al (2008) GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. Nucleic Acids Res 36 (Database issue):D1034–D1040

McKay SJ, Vergara IA, Stajich JE (2010) Using the Generic Synteny Browser (GBrowse_syn). Curr Protoc Bioinformatics, Chapter 9: Unit 9.12

Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G et al (2009) BioMart—biological queries made easy. BMC Genomics 10:22

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A et al (2002) The generic genome browser: a building block for a model organism system database. Genome Res 12(10):1599–1610

Wegrzyn JL, Lee JM, Tearse BR, Neale DB (2008) TreeGenes: a forest tree genome database. Int J Plant Genomics 2008:412875

Wegrzyn JL, Lee JM, Liechty J, Neale DB (2009) PineSAP—sequence alignment and SNP identification pipeline. Bioinformatics 25 (19):2609–2610

Wu CC, Huang HC, Juan HF, Chen ST (2004) GeneNetwork: an interactive tool for reconstruction of genetic networks using microarray data. Bioinformatics 20(18):3691–3693

Youens-Clark K, Faga B, Yap IV, Stein L, Ware D (2009) CMap 1.01: a comparative mapping application for the Internet. Bioinformatics 25 (22):3040–3042

Zhang H, Morrison MA, Dewan A, Adams S, Andreoli M, Huynh N et al (2008) The NEI/NCBI dbGAP database: genotypes and haplotypes that may specifically predispose to risk of neovascular age-related macular degeneration. BMC Med Genet 9:51